Norms and Reliability

Introduction to Classical Test Theory (CCT)

- **CCT** is a model for understanding measurement
- CCT is based on the ***True Score Model***
- …for **each person**, their **observed score** on a test is comprised of:
  - Observed score (X) = True score (T) + Error (E)
- …for **a population**, the **total variability/variance** in scores is comprised of:
  - Total variance (σ^2) = True variance ($\sigma^2_{t)}$ + Error variance (σ^2_e)

The concept of reliability is based on what's known as classical test theory. Classical Test Theory is a model for understanding measurement, and it's based on what is known as the True Score Model. For the True Score Model, the assumption is that when you do a test, and you get some test score, that's known as the observed score, for a single person, it's comprised of two components. Their observed score is comprised of some true score that they have, that is some true level of the trait or construct that you're trying to measure, plus some amount of error. That's for a single person. For a population, that is when you give your test to a large number of people, the total variability or variance that you observed in scores is comprised of what's known as true variability plus error variability. The same concepts apply to one single observed score and a group or population of scores. When we use the term variance, we're talking about a population of scores, that is how your test performs in a population of people, and for a single person, it's about one single observed score. But in both cases, that observed score or that total amount of variability that we actually obtain from our testing situation, is comprised of two components based on the True Score Model. That is that the observed scores or the total variance is comprised of a true score or true score variance plus some amount of error.

- **True score** is a person's actual true ability level (i.e., measured without error)
- **Error** is component of observed score unrelated to test-takers true ability or trait being measured
- **True variance** and **error variance** thus refer to the variability in a collection/population of test scores

A true score is a kind of theoretical score, that is that it reflects the person's actual, real, true level of ability on the construct or thing that you're trying to measure. That is, it's measured without any error. It's the person's actual, theoretical, true performance across that particular domain, or trait, or construct that's being measured. Error is a component of the observed score, that is it's unrelated to the test-takers true ability or trait being measured. The true score is a theoretical quantity. We can never actually know a person's true level of ability on a construct or trait being measured, because every time that we try to measure it, there's going to be measurement error. That measurement error could be very small, or it could very large, and that is what we're trying to minimise when we're trying to our measurement. But in theory, we can never know a person's actual true score, because when we try and measure that, in any way, shape, or form, we're going to have some kind of error in our measurement, which is going to create noise in that data. But that's what the Classical Test Theory is based on. The notion that there is a true score and every time we try to measure it, there is some level of error, and you need to be able to account or understand the kind of error that is occurring whenever you do a measurement. That was referring to a single observation, and when we're talking about the population of people doing the test, we're talking about variability, so in other words, we're talking about true variance, that is the variability in scores, that is due to some differences between the true scores of the different people versus variance that is caused by variability in the error that is occurring in each of those measurements, across the population of people you're examining.

Reliability refers to consistency in measurement

- According to CCT: reliability is the proportion of the total variance attributed to true variance
- Reviewing our formulas, if:
  Total variance (σ^2) = True variance (σ^2_t) + Error variance (σ^2_e)

then

Reliability = True variance (σ^2_t) / Total variance (σ^2) = True variance (σ^2_t) / True variance (σ^2_t) + Error variance (σ^2_e)

Reliability, from a theoretical standpoint is really about consistency in measurement. That is, that you have some measurement tool, and it is consistent in how it's assessing or measuring or quantifying the construct of interest in any way that you're trying to measure it. So it could be consistency across items, it could be consistency over time, it could be consistency in any aspect of the test. If you're trying to examine whether there is reliability, you're really looking at consistency. According to classical test theory and True Score Model, reliability has a very specific definition. That is, that the total amount of variability that we observe in our test, in other words, if we gave to a population of people, the total amount of variability or the total variance is comprised of true score plus error variance. Based on True Score Model and Classical Test Theory, reliability is simply the proportion of the total amount of variation in our test, that is due to true score variance. You have true score variance over the total variance, which we know the total variance is comprised of true score variance plus some amount of error variance. That's what reliability is. When we talk about reliability, or ways to examine reliability in our test, we're trying to estimate that proportion, somehow. We're using different mathematical formulas to try and estimate the proportion of the total variance that is attributable to true score variance, and there are different ways to assess reliability, but they're all trying to get at that aspect of it. That is, the proportion of the total variance that is attributable to the true score variance.

Reliability Estimates

Reliability = True variance (σ^2_t) / Total variance (σ^2) = True variance (σ^2_t) / True variance (σ^2_t) + Error variance (σ^2_e)

What would happen to the reliability estimate if your test has a lot of error?

	Reliable test (error is low)	Unreliable test (error is high)
Total variance	100	100
True variance	90	30
(Error variance)	(100 – 90 = 10)	(100 – 30 = 70)
Reliability coefficient	90/100 = 0.9	30/100 = 0.3

If the proportion of the total variance was 0.9, that is, the true score variance was 90 and the total variance was 100, then our error variance would be implied by that as being 10, and our reliability coefficient, based on our formula would be 0.9. So, that would be a highly reliable test. That is, there's a large proportion of the total variance that's attributable to the true score variance. By contrast, we have another example, where we have high error, that is our true score variance, or the proportion of the total variance attributable to true score variance, is fairly low, only 30. Our error variance would then be high, 100 minus 70, and that would give a reliability coefficient of 0.30. The reliability coefficient is directly influenced by the proportion of true variance in the data. We don't know the true variance, and we can never know, because it's a theoretical, implied number, because every time we try to measure our true variance or the true score on a particular test, we're going to have measurement error.

What is error?

***Measurement Error*:**

- **Systematic Error** (source of error that is constant, proportionate, predictable)
- **Random Error** (source of error that is unpredictable, inconsistent, unrelated, i.e., noise)

When we're referring to error in our formulas for reliability, we're talking about measurement error. There are other sources of error, but we're talking about measurement error. Measurement error in our testing, is comprised of both systematic error and random error. Systematic error is generally a constant, proportionate, or predictable amount of error. If you knew what the source of the systematic error was, or you knew how much it influenced your score, or true variance, then it would be the same for everyone. In that way, it's not predictable in the sense that we know what the variable is, or what the influence or the source of systematic error is, but it's predictable in how it affects our measurement. An example might be if we know that being in a noisy classroom affects everyone the same way, so they're trying to do some kind of test, and being in a noisy classroom made everyone perform 10 points worse on the test, compared to some other classroom that had a very quiet environment, we would say that's a source of systematic error, because everyone had the same or constant or predictable effect on their observed score, and so if we know that systematic error or the amount of that systematic error, we could adjust for it. That is, as a scientist, we could say that we're aware of this source of systematic error, we know how much it will be affect our observed scores, and so we can account for that through adjustment of our observed score. By contrast, random error is unpredictable, inconsistent, unrelated to the person's true score or the environment, or things like that. Random error is the good error, because on average, for some people, their error will cause their observed score to be higher than their actual true score, whilst for some people, their observed score will be lower than their true score due to this error. But on average, we would expect in a population that the random error will mean that the mean score for the test will be approximately what the mean would be for the population. That doesn't mean that for any single person, their observed score will be a good reflection of their true score. It just means that because it's unpredictable, and there's not much you can do about random error, you just want to make sure that you can try and have an understanding of how much random error there is in your data, to then be able to make adjustments about the precision that you apply into the different observed scores that you obtain. One thing worth noting is that when we're talking about reliability, even though our error variance is comprised of systematic error and random error, we're really only trying to measure random error. That's because reliability is always about consistency in measurement. In other words, if you have a systematic source of error that is constant, proportionate, predictably affecting your observed score, then that means it's probably a reliable source of error, and so any observed scores that we obtain will all be shifted in some systematic way, and so when we're looking at our reliability estimates or we're calculating our reliability, we're really trying to get a feel for how much random error there is in our data, and we are actually unable to have a good understanding of systematic error based on our reliability formulas. So, systematic error is something that you need to think about from a theoretical standpoint, from a standpoint of understanding the situations in which people are taking tests, and from that, you can then try and deduce what's going on with systematic error. Another way to deal with systematic error or to have a better understanding of it is through examining validity. If you find that in your classroom, your class scores consistently 10 points higher or lower than some other group, it will provide you an opportunity to have a think about what it could be about your classroom, in particular, that may be affecting the observed scores, or causing them to have systematic error.

Sources of Measurement Error

Test Construction:

Variation due to **differences in items on same test or between tests (i.e., item/content sampling)**.

Test Administration: variation due to testing environment

- ***Test-taker variables*** (e.g., arousal, stress, physical discomfort, lack of sleep, drugs, medication)
- ***Examiner variables*** (e.g., physical appearance, demeanour)

Test Scoring and Interpretation:

Variation due to differences in scoring and interpretation.

Sampling Error:

Variation due to representativeness of sample.

Methodological errors:

Variation due to poor training, unstandardised administration, unclear questions, biased questions.

The first source of measurement error is related to the construction of your test, or test construction. That is just talking about how error variation can creep into your test, because you just have different items, they're worded differently, and even though they're trying to get at the same construct, there's just variability between the items that makes them different and cause some amount of error. Another source of measurement error is related to test administration, so that is due to the testing environment, so that could be related to the person actually taking the test, how aroused they are, if they're stressed out, if they're in some kind of noisy environment and things like that, and how that affects that person, or it could be related to the examiner themselves, it could be that they seem scary or their physical demeaner is affecting the test performance, but it's related to the environment in which the person is taking the test. Another source of measurement error is related to the scoring and interpretation of the test. It could be that different people examine items on a test and score them in different ways. It's very common in IQ tests that some of the items that you have are qualitatively assessed, that is there's no single correct score, and so there is some variability that might creep into your test, some error variance, because different people that are scoring the test may interpret what's going on a little bit differently, and so that can be a source of measurement error. Sampling error is a little bit of an unusual one, because sampling error is something that can be addressed at the population level. So for any one test score, there's not really sampling error, because you have that one person, that one person is the only sample of interest, that is that obviously if you've got a population of one person, then there's no sampling error, but for the mean scores on your test, for say not some normative data or something like that, the larger the sample that you have, the smaller your sampling error. Sampling error is one of those sources of error that can creep into your measurement tool and can be addressed just by having a larger sample, and if you have the entire population there, then of course we'll have no sampling error at all. Finally, we have methodological errors, which again, is mostly around the design of the test and the way in which it's administered by the test administrator, so if there's no systematic instructions or the questions to the participants are unclear, a lot of those factors can cause you to have error in your data, and cause differences between people to be due to error rather than true score variance. In summary, we have a number of sources of measurement error, all of which you need to consider when examining the psychometric properties of your test.

CCT True Score Model vs Alternatives

- True Score Model of measurement (based on CCT) is simple, intuitive, and thus widely used
- Another widely used model of measurement is **Item Response Theory (IRT)**
- CCT assumptions more readily met than IRT, and assumes only two components to measurement
- But CCT assumes all item on a test have an equal ability to measure the underlying construct of interest

The reason why we use the True Score Model and Classical Test Theory is that it's fairly simple, intuitive, and because of this, it's very widely used. But it's not the only model of measurement, and another source, or another model of measurement that is increasingly being used is known as Item Response Theory. Item Response Theory is a bit more of a complex model, but it is increasingly being used because the statistics that underly Item Response Theory are becoming increasingly more available in software and things like that. It's increasingly being easier to implement and use to examine the psychometric properties of your items or your test. It's important to note that in Classical Test Theory, the assumptions are more readily met than IRT models, and it only assumes that any observed score is

made up of two components, that is, your true score plus some amount of error. But that might not be the best model for what is occurring in any observed score. The main reason is that Classical Test Theory assumes that all items on a test have an equal ability to measure the underlying construct of interest. For example, let's say we have a test that is trying to measure risk-taking, and let's just say we have two items, one of which asks have you ever received a speeding fine, the implication being that if you've received a speeding fine, you probably are some kind of risk-taker. Then we have another question which says have you used heroin before, or have you injected heroin. Now obviously those two items are very different, even though in our risk-taking measurement tool, they're probably both related to risk-taking, but you can see that one of those items, getting a speeding fine, is probably much more commonly experienced by the population than something like using or injecting heroin. So, according to Classical Test Theory, those two items would be equivalent in their ability to measure risk-taking, but that's probably a fair assumption. It's probably that those two items perform very differently or are related to risk-taking in very different ways. It could be that a lot of people who have speeding fines might be risk-takers, but also, it could be that they're only mild or moderate risk-takers, whereas if you're using heroin, you're probably at the high end of risk-taking. Item Response Theory is one of those models of measurement that can kind of step into and address some of the limitations of Classical Test Theory, by examining items specifically as they relate to a particular construct of interest and understand how those items perform differently in relating to the underlying trait or construct that is attempting to be measured.

Item Response Theory (IRT)

- IRT provides a way to model the probability that a person with *X* ability level will correctly answer a question that is "tuned" to that ability level
- IRT incorporates considerations of item **Difficulty** and **Discrimination.**
  - *Difficulty* relates to an item not being easily accomplished, solved, or comprehended
  - *Discrimination* refers to the degree to which an item differentiates among people with higher or lower levels of the trait, ability, or construct being measured.

IRT provides a way to model the probability that a person with some level of ability, so in the True Score Model, it would be some true score, will correctly answer a question that is tuned to that level of ability. This is essentially a way in which you can say if you actually truly score some level of risk-taking that is in the moderate range, this item is really good at tweaking or tuning into that change between people that occurs at that level of the latent trait. The item that relates to how many times have you been speeding, if it's related to the low end of risk-taking, it would be a really good item to tune out differences at the low end of the trait, whereas if the heroin item, asking someone whether they've used heroin, is probably much better at tuning differences between people at risk-taking levels at the higher end of risk-taking, because at the lower end of risk-taking, no one is going to be answering that they're using heroin. It's only going to be tuning in to different levels of the latent trait being examined, and that's what Item Response Theory can do. It can tell you how your items are tuning in to different levels of the latent trait being examined. At the basic level, IRT models incorporate what's known as item difficulty and discrimination parameters or measures of what's known as item difficulty or item discrimination. Item difficulty relates to an item not being easily accomplished, solved, or comprehended. That is, how difficult is the item to answer yes to, so an item like the heroin item, do you use heroin, would have high difficulty, that is that you would really need to be high on the latent trait of risk-taking in order to answer that question positively. It's not about the difficulty in solving some mathematical problem or anything like that, to answer that question. It's that you need to have high levels of the trait being examined in order to positively endorse that particular item, and in IRT models that is called item difficulty. Item discrimination, on the other hand, refers to the degree at which an item differentiates people with higher or lower levels of the trait. If we think about our item related to speeding, and we say how many times have you received a speeding fine, we would probably expect a very broad range of risk-takers to endorse that item. It's not like it's some level of risk-taking where all of sudden people start getting speeding fines. That item would have low discrimination, because at a very wide range of the trait, that item is something that's going to be endorsed positively. Having high item discrimination is probably not something that you want. You want to have items that discriminate very clearly amongst different levels of the latent trait.

Reliability Estimates

Because person's true score is unknown, we use different mathematical methods to **estimate** the reliability of tests.

Common examples include:

- Test-retest reliability
- Parallel and alternate forms reliability
- Internal consistency reliability
    - E.g., split half, inter-item correlation, Cronbach's alpha
- Interrater/interscorer reliability

Because we can never know a person's true score, we use different mathematical methods to estimate the reliability of our tests. There are many different ways to estimate reliability, some of common ones include test-retest reliability, parallel and alternate forms reliability, internal consistency reliability, and also interrater or inter-scorer reliability.

Test-retest reliability

Test-rest reliability is an estimate of **reliability over time**.

- Obtained by correlating pairs of scores from same people on administration of same test at different times
- Appropriate for stable variables (e.g. personality)
- Estimates tend to decrease as time passes

Test-retest reliability is an estimate of reliability over time, so it's about the consistency of your test or the consistency in scores from your test, over some amount of time, which could be one day, one hour, one year or one decade. It's about how reliable, how consistent are your observed scores over time, and you obtain that by correlating pairs of scores from the same person, administered at different times. Test-retest reliability is appropriate for variables that are assumed to be stable, something like personality. It is not appropriate to use test-retest reliability for something that is supposed to be fluctuating like mood, so there's no point examining test-retest reliability if you expect that your mood could change over the short period of time that you're going to be administering your retest. And again, it's about the stability and the timeframes, so if you think that something could be stable over one week, well it's fine to do a test-retest reliability over one week, but over ten years, you would expect there to be some change, and so you obviously won't do test-retest reliability over the ten year period. Just keep it to the period in which you're going to expect it to be stable. On average, over time, even the most stable variables will expect to have a lower test-retest reliability. Estimates tend to decrease as time passes.

Parallel and Alternate Forms Reliability

- **Parallel forms**: two versions of a test are parallel if in both versions, the means and variances of test scores are equal
- **Alternate forms**: there is an attempt to create two forms of a test, but they do not meet strict requirements of parallel forms
- Obtained by correlating the scores of the same people measured with the different forms

In parallel and alternate forms reliability, they're very similar in that you are trying to create two versions of your test, and then trying to examine whether the two versions perform the same way. There's consistency in scores across the different versions. Commonly in psychology, we might want to see if a person's cognitive functioning might improve over a short period of time, so it might be that a person comes in and has some kind of head injury, so we give them a short test to see how their cognitive functioning is going, but over a short period of time, like a couple of days, we might want to see how their functioning has improved or gotten worse. We would want to make sure that we have the same test, the same kinds of items, but different versions. We can't give them the exact same version because they might remember the answers that they provided previously. We want to have the same test but different versions, and we

want those versions to be equal. If they're not equal, then any differences that we observe between the versions could be due to error. We want make sure that our test is reliable, consistent, across the versions of the test. In parallel forms of the test, parallel forms of a test is a mathematical question. That is, when we create our two forms and we obtain some data from the same people on the two different forms of our test, if the mean and variances of the test scores are equal, then we can say that the tests are parallel. If the means and variances for the two versions of the test are not equal, then we can say that the tests are just alternate forms, and we would assume though that at least there is some correlation between the two tests. It's a mathematical difference, parallel is a more strict form of alternate forms. Ideally, we want parallel tests but often that's not possible.

Internal Consistency

Split-half reliability: obtained by correlating two pairs of scores obtained from equivalent halves of a single test administered once.

Entails three steps:

- Step 1. Divide the test into two halves.
- Step 2. Correlate scores on the two halves of the test.
- Step 3. Generalise the half-test reliability to the full-test reliability using the Spearman-Brown formula.

Internal consistency relates to whether there's consistency in how participants will answer items on your test, and there are a number of ways to examine internal consistency. Some of these are probably not used widely today or used anymore, while some of them need to be used with caution. One form of internal consistency is known as split-half reliability, and that's obtained by correlating two pairs of scores obtained from equivalent halves of a single test administered once. So, the steps to obtain split-half reliability are three-fold. The first thing you need to do is, after you administer your test, you divide it into two, and how you divide it into two is up to you. Maybe you take the odd and even items, maybe you split it in half, it's really up to you how you do that, you want to try and make that as random as possible. Then you correlate the scores on the two halves of your test, but then there's a third step, because you can't use the correlation between the two halves of your test as a good representation of reliability, because that correlation is based on only half the items in your full test. So, you use what's known as the Spearman-Brown formula to generalise the half test reliability, so that's the correlation between the two halves of your test, to what the reliability would be if you had used all the items.

Spearman-Brown (S-B) formula

S-B formula allows one to estimate internal consistency reliability from a correlation between two halves of the one test.

It is given by: $r_{sb} = nr_{xy} / 1 + (n - 1)\, r_{xy}$

- r_{sb} = SpearmanBrown predicted reliability
- n = #items on new version / #items on current version
- r_{xy} = correlation between half x and half y

The Spearman-Brown formula allows you to estimate the internal consistency reliability from a correlation between two halves of the one test, and we want to do that because reliability in general will be increased when you have more items. In other words, if you have a ten-item test and a hundred-item test, being able to triangulate the true score variance or the true score of a person will be higher the more ways you can interrogate that latent variable or construct being examined. On average, in general, having more items on a test will increase your reliability, so if you just use the reliability or the correlation between the two halves of your test, you're going to be doing yourself a disservice, because your reliability is probably going to be higher should you have used all of the items on your test. The Spearman-Brown predicted reliability is equal to n times the correlation between the two halves of your test, divided by 1 plus n – 1

times the correlation between the two halves of your test. You'll notice that n is the part that changes depending on how many items you had on your test, and how many items used in the split-half correlation. N is the number of items on the new version divided by the number of items on the current version. So, if you use a split-half reliability, the items on the new version would be the number of items you expect to have in your full test, and the number of items on the current version is the number of items you used in the split-half correlation.

Example: Let's say I have a 30-item test (i.e., split-half = 2x15 items) and the split-half correlation (r_{xy}) was = .70

Using the S-B to estimate reliability	Original 30-item version	Shortened 5-item version	Lengthened 100-item version
r_{xy}	.70	.70	.70
n	30/15 = 2	5/15 = .33	100/15 = 6.66
R_{sb}	2(.70) / 1 + (2-1) (.70)	.33(.70)/ 1 + (33 – 1) (.70) = .44	6.66(.70)/ 1 + (6.66 – 1) (.70) = .94

Internal Consistency

Other Methods of Estimating Internal Consistency

- **Inter-item consistency/correlation**: the degree of relatedness of items on attest. Able to gauge the homogeneity of a test.
- **Kuder-Richardson formula 20**: statistic of choice for determining the inter-item consistency of dichotomous items.
- **Coefficient alpha**: mean of all possible split-half correlations, corrected by the Spearman-Brown formula. The most popular approach for internal consistency. Values range from 0 to 1.

Other measures of internal consistency are simply getting the average of your inter-item correlations, so you have a 5-item measure, you just get the correlations between all 5 items and get the average of those, and that would be your average inter-item consistency, or inter-item correlation. This can give you some idea about the homogeneity or internal consistency of your test. Another way of examining internal consistency when you have items that are binary-coded or dichotomous is the Kuder-Richardson formula, and coefficient alpha. Cronbach's alpha is mathematically the mean of all the possible split half correlations that's been corrected by the Spearman-Brown formula. A lot of the time when we have mathematical equations, you can show that it is equivalent to some other thing, and so, coefficient alpha just happens to be equivalent to if you had done all the possible split-half correlations in your dataset. This means that you can use a simple formula to obtain a number that would be difficult to calculate by hand. Cronbach's alpha is a very common measure of internal consistency, and probably the most commonly used measure in psychology, although it does have some problems.

Care with Cronbach Alpha

- Cronbach alpha is often incorrectly used
- Lower estimate of reliability
- Not a measure of unidimensionality
    - i.e., it is a function only of the number of items, and the average inter-item correlation

Even though Cronbach's alpha is probably the most commonly used measure of, not even just internal consistency, but reliability in general, in psychology, it's important to be aware of some of the limitations of Cronbach's alpha, because it's not the be-all and end-all or gold standard, it has a number of faults. It's often incorrectly used, it's often applied without really thinking about the items or how they are related to each other, or whether the items are intended to be homogenous or not. You should only be looking at internal consistency and Cronbach's alpha if you think the items on your test are indeed homogenous. If you have items that are measuring

different things, that is they're not homogenous, then there's no way you should be using Cronbach's alpha. You should only be using it for items that you theoretically or have attempted to develop so that they are measuring the same thing, or they're being responded to or answered by the participants in the same way. It's a lower estimate of reliability. There are other measures of internal consistency, like coefficient omega and other variations that are theoretically supposed to be better representations of the true reliability, with Cronbach's alpha being what is essentially a lower estimate of reliability, so it's more cautious in that sense. It's not a measure of unidimensionality. Internal consistency is about whether the items on your test are being responded to in the same way. That's what we would call homogenous items. The items are the same kind of thing, they're being responded to in the same way. That doesn't mean that the items are tapping into only one dimension or one latent variable. Let's say we have an item called how many times have you received a speeding fine versus how many times have you injected heroin. Both of those items are probably going to be answered consistently, in other words, if you're going to engage in heroin use, you're probably also going to be speeding because of its relationship with speeding. But it might be that in risk-taking, those two items are more specifically related to other specific dimensions within the broader dimension of risk-taking. It could be that speeding-related risk-taking is it's own little dimension, and drug-related risk-taking is its own little dimension as well. So, even though those items are probably going to be answered in the same way, in other words, they're going to be homogenous, it doesn't necessarily mean that they belong to one underlying construct. It could be multiple constructs, and the only way you can determine the dimensionality of your items is to use factor analysis. So, it's not a measure of unidimensionality. It is only a measure of internal consistency, and so it's only a function of the number of items, and the average inter-item correlation, based on the Cronbach's alpha formula. It's only based on those two things, and you would need to use something like factor analysis to really interrogate the dimensionality or how many underlying constructs, how many latent variables, underly the items that you're using in your measure.

Variable	1.	2.	3.	4.	5.	6.	
1.	-						
2.	.8	-					
3.	.8	.8	-				
4.	.3	.3	.3	-			
5.	.3	.3	.3	.8	-		
6.	.3	.3	.3	.8	.8	-	

Cronbach a = .86

Variable	1.	2.	3.	4.	5.	6.
1.	-					
2.	.5	-				
3.	.5		-			
4.	.5	.5		-		
5.	.5	.5	.5		-	
6.	.5	.5	.5	.5	.5	-

Cronbach a = .86

Interrater/interscorer

Interrater/Inter-scorer reliability is degree of agreement/consistency between two or more scorers (or judges or raters).

- Often used with behavioural measures
- Guards against biases or idiosyncrasies in scoring
- Obtained by correlating scores from different raters:
  - Use intraclass correlation for continuous measures
  - Use Cohen's Kappa for categorical measures

Interrater or inter-scorer reliability is basically the degree of agreement or consistency between two or more scorers, or judgers, or raters of a test. It's often used with behavioural measures. The classical example is examining what's known as the strange situation paradigm, which is a measure of attachment in children and what you observe, which you have to videorecord are the interactions that the child has with their mother, and their behaviour when their mum leaves the room, and leaves the child with a stranger. That's a videorecorded behavioural observation, and you need to score it, and make sure that there is pretty good interrater or inter-scorer reliability. In other words, the scoring that is obtained for each participant that you see is not dependent on the particular scorer, but there is some reliability amongst the different judges or raters of those behaviours. It guards against biases or idiosyncrasies in scoring. Like most of the other measures of reliability, it involves a correlation, and that is correlating the scores obtained or ratings obtained from different raters or scorers or judges. If you have continuous measures, you want to use an intraclass correlation, and if you have categorical measures you want to use Cohen's Kappa. Those two measures allow you to examine whether there are systematic differences between the two. So, it could be that person A rates 10 points higher than person B, then that correlation won't account for the fact that there is that systematic difference. Something like an intraclass correlation will adjust the correlation coefficient to make sure that there isn't any systematic difference between two raters or scorers of a test.

Choosing Reliability Estimates

The nature of the test will often determine the reliability metric, e.g.

- Are the test items homogenous or heterogenous in nature
- Is the characteristic, ability, trait being measured presumed to be dynamic or static
- The range of test scores is or is not restricted
- The test is a speed or a power test
- The test is or is not criterion-referenced

Otherwise, you can select whatever you think is appropriate.

We need to use an estimate of reliability because we can never know the person's true score, and therefore we can never know how much error there is the test. But we can estimate the reliability based on giving our test to a sample of people, and then obtaining whatever reliability estimate that we want to use. Which reliability estimate you should use is really dependent on a number of factors, but all things being equal, it's really on personal preference. What's meant by all things being equal is that if there are a number of options, for example, for internal consistency, that you can choose, so long as the option is appropriate for the nature of your test, it doesn't matter which one you pick. It's really about what is perhaps more commonly used in your field and might even be about what's easier to do. But in terms of the nature and the item scaling, there are a number of factors that you need to consider. Firstly, is whether you think the items are homogenous or heterogenous, in other words, you shouldn't use something like Cronbach's alpha if you think that there is heterogeneity in the items, or the items are not intended to be answered in a consistent way. You want to consider whether the thing that you're trying to measure is assumed to be dynamic or static, so if you think about test-retest reliability, where there's no point using test-retest reliability if you think there is some dynamic nature to the thing that you're measuring. You want to look at whether the range of your test scores is or is not restricted. What's meant by this is say that you have a five-item scale, and everyone does really well, with everyone getting a 3 or a 4 or a 5, and so, when you're correlating items where everyone scores pretty much the same way, you're going to have a very restricted range of variation in your data, and so that will affect any correlations you observe. The same would be true if your scale range from say 0 to 1000, where any correlations would be inflated, and this is a well-known fact of a correlation coefficient, that restricting the range will attenuate correlations and increasing the range will inflate those correlations. So, it's important to consider how much variation you are using or how much variation you have in your items, how restricted the range of test scores when considering what's going to be an appropriate measure of reliability, and how to interpret that reliability in light of the variation of your test scores. Another thing to consider is whether your test is a speed or a power test. A speed test is when you say, I've got a 100 items, they're all the same, I'm going to give you 2 minutes, and I want to see how many you can do, and so the idea of a speed test is that all those items will be

approximately equal in difficulty, and it's just really about how fast someone can be at completing those items. A power test, by contrast, means that you have increasing difficulty over the items. So, if you think about a maths test, the earlier items are probably going to be much easier than the items towards the end, and those ones towards the end are what kind of separates the higher and the lower performing people on that test. So, you can imagine that a power test, having items of different levels of difficulty, it doesn't make sense to use some measures of reliability, like it wouldn't be useful to use a measure of split-half internal consistency, if half of your test is going to be much more difficult than the other half. By contrast, a measure of internal consistency for a speed test, where you want all the items to be approximately answered the same way, would probably be a really good measure of reliability for that type of test. So, it's important to consider whether your test is a speed or a power test when considering which reliability estimate to use. Finally, it's important to think about whether your test is criterion-referenced. A criterion-referenced test is a test that in order to pass you need to reach a certain threshold, so once you pass, you pass. And if you reach the end of the threshold, you have passed the test. So, looking at reliability in that context, like an internal consistency reliability may not make sense. But all things being equal, if you've considered those factors and you think that you still have different options for which reliability estimate to choose, then it's up to you which one you choose. So long as you understand the limitations of the reliability estimate that you are selecting, it's really about personal preference and probably what is the most commonly used measures in your field of study.

How do we account for reliability in a single score?

- Our reliability coefficient tells us about error in our test **in general**
- We can use this reliability estimate to understand how confident we can be in a single **observed score for one person**

Our reliability estimates have been talking about reliability of our test in general, and to obtain our reliability estimates we've generally been applying or administering our test to a population of people, and then used the data from that population of people to get an estimate of the reliability. However, that doesn't tell us anything really about a single person, so a single person who performs our test, and if you think about a clinical setting, that's what we've got. We've got one person performing our test, and what we want to be able to do is understand that single person's observed score and use the reliability that we have estimated for the total test to then have a better understanding of the precision that any single observed score will have. That is, we want to be able to look a single person's score, and then have some level or some ability to look at that score and understand how close we think this is to the person's true score, after we account for the reliability in our measurement tool.

The Standard Error of Measurement (SEM)

SEM provides measure of precision of an observed test score i.e., estimate of amount of error in an observed score.

- Generally: higher reliability = lower SEM
- SEM can be used to estimate the extent of deviation between observed and true score

$$SEM = \alpha_{meas} = \alpha \sqrt{1 - r_{xx}}$$

α = standard deviation of sample

r_{xx} = reliability coefficient

95% Cl of test score = test score ± 1.96 (SEM)

The way we can use the reliability of our test to get a better understanding of how precise an observed score is, is to use what's known as the standard error of measurement. The SEM for short provides a measure of the precision of an observed score, or the estimate of

the amount of error in an observed score based on the reliability that we have estimated from our population of people. Generally, the higher your reliability of the test, the lower your standard error of measurement will be, and you can use the standard error of measurement to estimate the extent of the deviation between the observed score that you see and the true score. So, if you get a very small standard error of measurement, you can be fairly confident that the observed score that you are seeing is a very close representation of the person's true score, and that there isn't a lot of error in that observed score. The formula for the standard error of measurement is given by sigma times the square root of 1 minus the reliability of the test. Sigma is just the standard deviation of the sample, so if you think about this from a practical perspective, if you're looking at one single person's observed score, to get the standard error of measurement, you need to go to the sample that you've obtained, you need to get the standard deviation of your measurement tool in that sample, and then you get the reliability coefficient, and it doesn't matter which reliability coefficient you use, because the reliability coefficient you use could be different depending on your particular survey. You might use test-retest reliability, you might use Cronbach's alpha, you might use both of them, and get the average of them. But you just need to get some estimate of the reliability of the test and internal consistency reliability is as good as any other. If you plug that in there, you can see how you can use the sample that you used to obtain the reliability, to then get some estimate of the precision of any single observed score based on the standard error of measurement, and you can get a confidence interval around an observed score. So, the 95% confidence interval for a test score or an observed score is the same as any other confidence interval, so 95% confidence interval would be the observed score or test score plus or minus 1.96 times the standard error of measurement.

If you change the reliability of your test for any consistent standard deviation, the standard error of your measurement will decrease. That is, you will get more precision in your observed score. If your observed score was 100, you will see the confidence interval getting smaller around that observed score of 100, if your reliability increases and your standard deviation remains the same. Similarly, if you had a change in standard deviation, in other words, the variation that you observed in your data to begin with, for any level of reliability, you will see that the standard error of measurement will increase, so that the more variation that you have in your data, will cause your standard error of measurement to increase, because you're going to have more noise, you've got more variability in your data, and you're just naturally going to have more noise. But again, you can minimise the effects of having variation in your data by increasing your reliability. You will see that the high reliability of 0.9 compared to 0.5, those standard error of measurements are much smaller for the same level of standard deviation, and that corresponds to the changes in the width of the confidence intervals as well. It's important to note that the standard deviation of your test is not something that is easily manipulated. You give a test and you're just going to have variability. In other words, that's not the part that you can really manipulate unless you change the scaling of your items, or you change the difficulty of your items to make the data have more or less variability. But what you can change is the reliability. You can increase the precision by affecting your reliability, by dealing with standardised instructions, test administration, any environmental influences on the test-takers performance, the clarity of the items, or the way the test is constructed, all of these factors we have control over, and so, the best way to increase the confidence that you have that any single observed score is a true representation of the person's true score is to increase the reliability of your test.

Standard Error of the Difference (SED)

- The SED is measure of how large a difference in test scores would be to be considered "statistically significant"
- Helps with three questions (Note: test 1 & 2 must be on same scale):
    - How did Person A's performance on test 1 compare with own performance on test 2?
    - How did Person A's performance on test 1 compare with Person B's performance on test 1?
    - How did Person A's performance on test 1 compare with Person B's performance on test 2?

$$\alpha_{diff} = \sqrt{2 - r_{xx(test1)} - r_{xx(test2)}}$$

$r_{xx(test1)}$ = reliability coefficient of test 1

$r_{xx(test2)}$ = reliability coefficient of test 2

95% CI of test score = test score ± 1.96 (SEM)

Extending on the standard error of measurement is the standard error of the difference. The standard error of measurement was when you want to look at just a single observed score and have a good understanding of how close that observed score is probably going to be to the person's true score. The standard error of the difference looks at the difference between two observed scores and is a way to determine whether the difference between two observed scores is statistically significant. Some common scenarios that involve creating a different score and then examining whether there is a statistically significant difference between two particular observations, again this is not significantly different between the means of different groups, this is just two scores, two observed scores and what is the difference between those two scores. So, a classic example might be that if you get person A and you examine their score on test 1 and compare that to their performance on test 2, what about if you want to look at person A's performance on test 1 compared with person B's performance on test 1, or what about if we look at person A's performance on test 1 compared to person B's performance on test 2. These are all very common scenarios. It could also be person A's performance on test 1 at two different timepoints. Basically, you've got two observed scores and you want to see whether the difference that you're observing is greater than would be expected based on the reliability of your test, that's the standard error of the difference. The formula is just given by square root of 2 minus the reliability of test 1 minus the reliability of test 2, and you can get the confidence interval for that as well. These are very commonly used formulas, very simple to calculate once you know the reliability of your test. One thing to note is that when you're comparing scores using the standard error of difference, you must standardise your variables or use standard or the same scale. Otherwise, you are not going to be comparing apples with apples, you're going to be comparing two different scores on two different scales, and so you need to make sure they're on the same standard scale to compare, or you will be misled.

Norms

- **Norm-referenced** testing and assessment: allows one to derive meaning from a person's test score by comparing it to a reference group.
- **Norms** are the data that is obtained from the particular group of test-takers that are being used as the reference group.
- A **normative sample** is the reference group to which test-takers are compared.

Norm-referenced testing and assessment allows you to derive meaning from a person's test score, by comparing it to a reference group. Let's say you have a patient, they come in, and do a measure of depression and get a score of 5. If there's no context to what the number 5 means within that measurement tool, there's really no way to interpret that particular score. You need some way to contextualise a person's particular observed score, and the way we do that in norm-referenced testing is to use norms. So, norms are the data that's obtained from a particular group of test-takers that are being used as a reference group. We have a person's observed score and in order to understand a single person's observed score, we compare them to a reference group, from which we obtain our norms. The reference group is also known as the normative sample, so we use our normative sample to derive norms, and from the norms we can compare our single observed score to the norms obtained from the normative sample in order to help contextualise a particular observed score.

Norm-Referenced and Criterion-Referenced Tests

- **Norm-referenced** tests compare an individual's score to the norms obtained from a normative sample.
- **Criterion-referenced** tests compare an individual's score to a particular predetermined standard, criterion, level of performance/proficiency, or mastery (e.g., a driving exam).

When we're trying to develop some kind of comparison, you want to think about whether you want to do normative comparisons or criterion-referenced comparisons or use norm-referenced tests or criterion-referenced tests. Norm-referenced tests compare an individual's observed score to the norms obtained from a normative sample. By contrast, a criterion-referenced test compares an

individual's score to a particular predetermined standard or a criterion or level of performance proficiency to pass. We can think about a norm-referenced test as classically being those things where you think I want to know how this person is doing relative to other people like them. A classic example might be an IQ test, where you compare the person's performance to other people of their age and educational level. Criterion-referenced tests, though, have some minimum level of standard or some criterion by which you're judging or comparing the person's performance. So, if you think about a driving test, there's no normative data for a driving test. You can do the driving test, or you cannot. The criterion is that you're a satisfactory driver. Now, determining how to determine what that criterion is, is based on how you develop your test and what you think is going to be appropriate. But a criterion-referenced test are those tests where you say I expect a certain level of proficiency in order to pass the test, and you cannot pass the test if you don't meet that certain criterion of performance.

Sampling to Develop Norms

- **Standardisation** is the process of administering test to representative sample to establish norms.
- **Sampling** is the selection of an intended population for the test, that has at least one common, observable characteristic.
- **Stratified sampling** purposefully includes a representation of different subgroups of population.
- **Stratified-random sampling** is sampling design that ensures every member of population has an equal opportunity of being included in a sample.

When developing norms, it's very important to consider the group of people that you're going to use as your normative sample. It's very important to consider how you're going to sample into that normative sample, so that you can make sure that your normative sample or normative group is representative of the population of interest. So, standardisation is the process of administering a test to a representative sample to establish norms. Sampling is the selection of the intended population for the test, and they need to have at least one characteristic that is measurable or of interest, so sampling could simply be that that one characteristic is they're an Australian citizen or it could be that they're a young person or in school. It's the selection of the intended population for the test. Stratified sampling is when you purposefully include a representation of different subgroups of the population. So, often these are around demographic factors, you want to make that you have males and females, you want to make sure that you have a sample that has Aboriginal and Torres Strait Islanders, you want to make sure that your normative sample has a variety of socioeconomic statuses. So, stratified sampling is ensuring that your sample does have a representation of different subgroups within your population, often different subgroups that you think would influence the outcome. Finally, we have stratified-random sampling where the subgroups that you have in your sample are specifically sampled so that they have an equal probability or opportunity of being included in the sample. So, what that means is that you forcibly make your sample to have an equal number of people within different strata. So, let's consider rural versus metropolitan residency, so if you live in the country versus if you live in the city. If you just did normal stratified sampling, you would, just by random chance expect to see a greater proportion of people from metropolitan city centres in your sample just because there's more people who live in the city than there are that live in the country. So, that means that even though you've randomly selected people from the population, if you've just randomly done it, then populations that have a small prevalence are going to also have a small prevalence in your sample, so they may not really be represented in my particular sample or normative sample. But stratified-random sampling means that you forcibly make an equal number of people from the different strata. So, let's say you use where you live as the strata, and the two different places could be the city versus in the country, and let's say you had a normative sample size of 1000, if you wanted to make that equal, you would say I'm going to forcibly take 500 people from the city and 500 people from the country, even though on average, we would expect there to be many more people in the city, but that makes sure that you have now two equal balanced samples, balanced by the strata, so that you can use your normative data for a more accurate comparison group.

Purposive sample is arbitrarily selecting a sample believed to be representative of the population.

Incidental/convenience sampling is a sample that is convenient or available for use. May not be representative of the population.

- Generalisation of findings from convenience samples must be made with caution.

Some other, less robust sampling methods are purposive sampling where you arbitrarily select a sample believed to be representative of the population. So, it might not be that it actually is representative of the population, but you think that you're doing the right thing. And then you have incidental sampling that it's just getting whatever sample that you can get, and the limitation of that is obviously that the sample may not be representative of the population, and so you need to be cautious when you're comparing an observed score to a normative sample that was obtained using incidental or convenience sampling. It's worth noting that a lot of the times in psychology, that's what we do because it's not only convenient, its cheap. It costs a lot of money to get a really strong representative sample from the population, and so often a lot of the tests that we use are based on convenience samples, and so you need to be making sure that you're interpreting any comparisons to a normative sample, based on convenience sampling with caution.

Process of Developing Norms

Having obtained the normative sample:

1. Administer the test with standard test of instructions
2. Recommend a setting for test administration
3. Collect and analyse data
4. Summarise data using descriptive statistics including measures of central tendency and variability
5. Provide a detailed description of the standardisation and administration protocol

The process of developing norms after you've obtained your sample involves a number of steps. The first is to administer the test, the test that you're trying to develop, with a standard set of instructions. It's extremely important that you have some standardised administration instructions so that you can be sure that everyone received the test or saw the test in the same way, and that way you can be sure that any differences between people are just due to differences in ability and not due to ability in how the test was administered. You need to recommend a setting for the test administration, and that's all part of the standardised instructions. You then need to collect and analyse your data, get your descriptive statistics including measures of central tendency and variability, like means and variation, and you need to make sure that you provide a detail of your standardisation and administration protocol so that anyone else that wants to use your norms understands the context in which the test was administered, to make sure that they are also administering their test in the same context and can then compare.

Types of Norms

Percentiles: the percentage of people in the normative sample whose score was below a particular raw score.

- Percentiles popular because easily calculated and interpreted
- Problem: real differences between raw scores may be minimised near ends of distribution and exaggerated in middle of distribution.

Some of the common types of norms include percentiles, which are the percentage of people in a normative sample whose score was below a particular raw score. So, if someone says to you, you scored in the 80th percentile, it means that 80% of the normative sample scored lower than you. You scored higher than 80% of the normative sample. Percentiles are popular because they're easily calculated and interpreted; it just means that you scored higher than whatever percentile in a normative sample, but there are a couple of problems because real differences between raw scores may be minimised near the ends of the distribution and exaggerated in the middle of the distribution. An example would be, if you scored in the 50th versus 55th percentile, there's not that much of a difference between you and the other percent because there's a lot of people in that. So, 5% of the population is not going to have a big difference in their observed scores at that level. But a 5% difference when you're at the 99th percentile versus the 94th percentile is a huge difference in

performance; you are really pushing the ends, and so even though it's still a 5-percentile difference, the magnitude of the difference in performance is extremely different, and so you need to make sure that when you're comparing percentiles, that you're considering is this happening in the middle of the range or is this happening towards the tails.

Age norms: average performance of normative sample segmented by age.

Grade norms: average performance of normative sample segmented by grade.

Subgroup norms: a normative sample can be segmented by any of criteria initially used in selecting sample.

National norms: derived from normative sample that was nationally representative of the population.

National anchor norms: equivalency table for scores for two different tests. Allows common comparison.

Local norms: provide normative information with respect to the local population's performance on some test.

Age norms are where your norms are just separated by age. In other words, you have a population, and you get the normative sample based on segmenting your normative sample by age groups; so, you can compare a person's at age 10 to other people of about the same age. This is the same for grade norms or subgroup norms; they're just norms for which the normative sample has been segmented by some particular factor. So, for grade norms, the normative sample is stratified by grade, and for different subgroups, it could be stratified by socioeconomic status, it could be by depression status, and so on. These are just different ways of dividing up the normative sample so that you can a more refined understanding of a person's score by comparing them to other people of a similar type to them; it could be similar age, grade, or some other factor that you want to compare to. National norms are norms that are derived from a normative sample that was nationally representative of the population. So, often these are obtained from the Australian Bureau of Statistics or really robust strong sampling approaches that cost quite a lot of money. National anchor norms are basically an equivalency table for scores on two different tests and allows a common comparison, or in other words, it's just a normative comparison so that for example, it says that if you score a 40 on some test, will we know that that's approximately the same as a 30 on some other test, and it allows you to then compare scores on different tests because there's some normative sample that shows a score on one test is about the same as a different score on a different test. Local norms, as the name suggests, is just normative information with respect to a local population and so again, it could be that we want local norms for Victoria or New South Wales or even to suburbs, districts, or neighbourhoods. Local norms are referring to the locality of the sample and where they're located and the normative data that's stratified by that particular factor. These are all about getting better comparisons; if you have your normative data stratified or segmented by age, by grade, by different subgroups, by locality, whether it's local versus national, all it means is that you'll have a better comparison group that you can compare your single person in front of you and their observed score to a more representative comparison group. So, you have a person in front of you who might be 25 years of age, who completed year 12, and is from a low socioeconomic status, if your normative data is stratified by those factors, you can then have a better comparison to compare your person's observed scores to other people in a similar circumstance or situation to them.

The Normal Curve

The normal curve is a bell-shaped, smooth, mathematically defined curve that is highest at its centre. Perfectly symmetrical.

Are Under the Normal Curve

The normal curve can be conveniently divided into areas defined by units of standard deviations.

We use the properties of the standard normal distribution to understand normative data and understand where our particular patient or person or observed score that we're looking at relates to the normative population. We know that approximately 68% will lie within one

standard deviation of the mean, based on the normal distribution, 95% of the population will be within two standard deviations from the mean, and 99.7% of the population, within three standard deviations of the mean. If we, in our normative sample, have a mean of 100 and a standard deviation of 10, if you scored more than three standard deviations from that mean, such as 140, based on the normal distribution we would say that would be very unusual; it was a very unusual score in the normative population to score that high. If you were within one standard deviation of that mean; 68% was within one standard deviation so it's fairly common. So in other words, your observed score is not that different from what we would expect based on the normative sample because your mean was closer to the normative sample mean. And so we can work out how many standard deviation units your observed score is from the normative sample to determine how common or how usual or typical your observed score is compared to the normative sample.

Standard Scores

Standard score is a raw score converted from one scale to another that has a predefined scale (i.e., set mean and standard deviation).

Z-score: conversion of a raw score into a number indicating how many standard deviation units the raw score is below or above the mean [$z = X - \bar{X} / s$]

Where X = raw/observed score; $\bar{X}$ = mean of the normative sample; s = standard deviation of the normative sample.

T-scores: aka "fifty plus or minus ten scale" – scale has set mean = 50 and standard deviation = 10.

An easy way to compare your observed score to the normative population is to convert your observed score into a standard score which is just placing your raw score on some set scale that you've predetermined based on some set mean and standard deviation. So there are many different versions of standard scores. The most common is a z-score standardisation. What that involves is converting your raw score into a number that represents how many standard deviation units your score was from the normative sample mean. Another type of standard score is a t-score aka a fifty plus or minus ten scale which as the name suggests is a scale that has a mean of 50 with a standard deviation of 10. So, if you score a 60, that means you're one standard deviation above the mean on a t-score scale. There are many other standard scores such as IQ. IQ generally will have a standard scale where you have a mean of 100 and a standard deviation of 15. So again, it doesn't matter what the scale is, so long as it's predetermined and once you've got everyone on the same scale, it's easier to understand where the person lies relative to the normative population.

Culture and Inference

- In selecting a test for use, responsible test should research all available norms to check if norms are appropriate for use with your patient.
- When interpreting test results it helps to know about the culture and era of the test-taker.
- It is important to conduct culturally informed assessment.

When you're selecting a test, it's important for test users to research all the available norms to make sure that you are using a normative sample that is going to be representative of the particular person or patient that you're administering your test to. There's no point comparing a person with a low educational status, so maybe they only went up to year 8, to people who have education on a university level. So, you need to make sure you, as the test administrator have researched all the available norms that there are, to see if there are particular normative data that is going to be better aligned with the particular person that you're going to be administering your test to. It's very important because if you don't that, you're going to get misled by how different the patient's observed score is from the normative sample, if that normative sample is nothing like them and it's not a fair comparison. So it's important to be conducting a culturally informed assessment at all times, and that includes selecting norms that are appropriate for the particular person that is sitting in front of you.